DOCUMENT RESUME

ED 467 811 TM 034 353

AUTHOR	Schnikpe, Deborah L.; Scrams, David J.
TITLE	Exploring Issues of Test Taker Behavior: Insights Gained from Response-Time Analyses. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-98-09
PUB DATE	1999-03-00
NOTE	28p.
PUB TYPE	Information Analyses (070) -- Reports - Evaluative (142)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Item Response Theory; Law Schools; Psychometrics; *Responses; Scoring; Timed Tests

ABSTRACT

The unobtrusive recording of item response times is one of the many advantages offered by computerized test administration. This report is a broad review of psychometric literature on response times. The review is not exhaustive, but does provide a sample of work that has been done. The review is organized into seven sections: (1) scoring models; (2) speed-accuracy relationships; (3) strategy usage; (4) speededness; (5) pacing; (6) predicting finishing times/setting time limits; and (7) subgroup differences. There is a final section that makes recommendations for future research. Among the research that may be particularly useful will be studies focusing on cognitive approaches. Psychometric response time researchers may benefit from looking at studies of response time in cognitive psychology. (Contains 3 figures and 97 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 811

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. VASELECK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Exploring Issues of Test Taker Behavior: Insights Gained from Response-Time Analyses

Deborah L. Schnikpe and David J. Scrams
Law School Admission Council

Law School Admission Council
Computerized Testing Report 98-09
March 1999

LAW Services

A Publication of the Law School Admission Council

TM034353

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and ***LSACD*** are trademarks of the Law School Admission Council, Inc.

Copyright© 1999 by Law School Admission Council, Inc.

All rights reserved. This book may not be reproduced or transmitted, in whole or in part, by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 2

Introduction 2
Response Times 3

Review of Research and Development to Date 3
Scoring Models 4
Speed-Accuracy Relationships 7
Strategy Usage 8
Speededness 9
Pacing 13
Predicting Finishing Times/Setting Time Limits 14
Subgroup Differences 15

Recommendations for Future Research 15

References 16

Executive Summary

In 1995, the Law School Admission Council (LSAC) began a five-year plan to research the advisability and feasibility of administering the Law School Admission Test (LSAT) by computer. The research plan is summarized in the Computerized LSAT Research Agenda. The Research Agenda suggests that the unobtrusive recording of item response times is one of the many advantages offered by computerized test administration. For example, analyses of item response times may lead to innovative ways to address speededness issues and scheduling problems as well as methods for identifying and investigating test taker strategies. Additionally, response times may reveal aspects of test taker proficiency that are not represented by response selections, and they may serve as an important consideration for equity and fairness reviews.

The present work is a broad review of psychometric literature on response times. It is organized into seven major sections: scoring models, speed-accuracy relationships, strategy usage, speededness, pacing, predicting finishing times/setting time limits, and subgroup differences. These categories are neither exclusive nor exhaustive but they serve as a general organization of the literature to date. There is a final section focused on recommendations for future work.

Several theorists have offered scoring models that depend either wholly or partially on response-time measures. These models generally include assumptions about the expected response-time distributions, relationships between response speed and response accuracy, and the nature of items for which the models are appropriate. The assumed response-time distributions are generally consistent with empirical results, but rigorous model-fitting tests are rare. Similarly, the assumed relationship between speed and accuracy often takes a form that is difficult to test empirically under standard test administrations. Finally, the items most often used as the basis for response-time models are more similar to simple cognitive tasks (e.g., perceptual speed tasks) than to items that are found on large-scale standardized assessments such as the LSAT. The general conclusion is that although scoring models incorporating response time are not yet ready for operational use, considerable advances have been made along these lines. Unfortunately, no validity studies have been offered that address the utility of the resulting scores.

Many of the scoring models rely on an assumed relationship between speed and accuracy, but a distinction must be made between within- and across-test-taker speed-accuracy relationships. Cognitive psychologists have been primarily interested in within-test-taker relationships (i.e., the speed-accuracy tradeoff: test takers are able to increase their response speed at the expense of accuracy or vice versa). Psychometricians have been primarily interested in across-test-taker relationships (i.e., fast test takers may be either more or less accurate than their slower counterparts). Some scoring models incorporate within-test-taker assumptions and others incorporate across-test-taker assumptions. Empirical psychometric work on speed-accuracy relationships, however, has focused exclusively on across-test-taker relationships, and work along these lines tends to show that the relationship depends heavily on test context and content. Useful scoring models must be flexible enough to accommodate these content and context effects.

Research on the use of response times to identify strategy usage is compelling but requires a reasonable understanding of possible strategies and their relationship to response times. Most of this work has used simple perceptual tasks (e.g., spatial-visualization tasks) or relatively well-understood cognitive tasks (e.g., mixed-number subtraction). This work has also been limited to differentiating between two strategies. Developing the requisite understanding of strategy usage for the item types currently used on the LSAT would require considerable further research, and using response times to identify these strategies would require extending the methodologies to account for multiple strategies.

Work on the use of response times to identify strategies has spawned a promising line of research on speededness. Traditionally, speededness work has focused exclusively on the number (and distribution) of unanswered items. With the availability of item response times, researchers recognized their value as additional sources of speededness information. Theorists have introduced the concept of rapid-guessing behavior: speeded test takers may quicken their pace considerably and begin answering less accurately as time expires. Such

behavior is readily identifiable by examining patterns of response times along with accuracy. This research would have direct implications for operational use and requires little future research for implementation.

Discussions of rapid-guessing behavior have also brought attention to more general issues of test taker pacing. Although little work has been done in this area, researchers have begun to explore methodologies for identifying different pacing strategies. Preliminary work suggests that test takers engage in numerous pacing strategies. These include devoting considerable time to early items followed by either rapid guessing or failure to finish, responding rapidly to early items to allow more time for later items, maintaining constant speed throughout an examination, and allocating time according to item difficulty. Some interest has been focused on subgroup differences in pacing strategies, and work in this area suggests that some subgroups may tend to engage in less optimal pacing strategies than other subgroups.

Some researchers have attempted to use response times to tackle the practical issues related to setting time limits. This work has generally involved predicting finishing times, but researchers have experienced only limited success. The standard approach has been to estimate total time on the basis of item response times or to predict finishing times on the basis of statistical regression. Neither approach has been overly successful, and field tests have often resulted in changes to the time limits suggested by previous research.

Response times may also be useful in terms of equity and fairness considerations. Some researchers have investigated subgroup differences in response times, and although the results are mixed, there appear to be only slight subgroup differences in response times. Unfortunately, most researchers have examined only mean or median response times rather than distributions of response times. Future work using full response-time distributions will likely be more rigorous. Investigating these issues within the context of a response-time model may also be advantageous.

Response-time research is growing in popularity along with the computerization of test administrations. Continued growth along the lines discussed in the present work is likely. There are also other important applications of response-time research that have received little attention. A close connection to cognitive approaches may be particularly valuable. Response time has been the preferred dependent variable in cognitive psychology since its inception, and psychometric response-time researchers may gain considerably by considering the results from this more established tradition. As LSAC considers the computerization of the LSAT, issues related to the use of response times should be of continued interest.

Abstract

The unobtrusive recording of item response times is one of the many advantages offered by computerized test administration. For example, analyses of item response times may lead to innovative ways to address speededness issues and scheduling problems, as well as methods for identifying and investigating test taker strategies. Additionally, response times may reveal aspects of test taker proficiency that are not represented by response selections, and they may serve as an important consideration for equity and fairness reviews. Response-time research is growing in popularity along with the computerization of test administrations. The present work is a broad review of psychometric literature on response times. The review is not exhaustive, but rather provides a sample of what has been done. A final section focuses on recommendations for future work.

Introduction

At its heart, assessment has the goal of inferring unobservable test taker characteristics from observable test taker behavior. Two obvious questions arise when we explore test taker behavior: "What inferences do we want to make," and "What observations do we use as the basis for these inferences?" A third question arises with respect to scoring, namely "How do we assign scores to test takers?" These three questions will be considered throughout this paper.

We would like to thank Lori McLeod, Peter Pashley, Chris Scarlata, and Kimberly Swygert for their helpful comments and suggestions on an earlier draft of this paper.

In both classical test theory (CTT) and item response theory (IRT), the focus is on accuracy and the inference being made involves test takers' overall proficiency in a domain of tasks (Mislevy, 1993). There are alternatives to focusing on accuracy and inferring proficiency. The bulk of this paper focuses on the uses of response times to infer various test taker characteristics. First, however, examples of work that uses neither response times nor strict accuracy will be discussed for the purpose of contrast with traditional accuracy-based approaches. The inferences being made in the following examples are test taker misconceptions, test taker skill profiles, and speededness.

The first example involves inferences about test taker misconceptions. In all of these instances, the overall score that a test taker receives is of secondary interest to the misconceptions that the test taker has in certain areas. Knowing what misconceptions the test taker has provides more detailed and useful information for remedying the misconceptions than a total score. Several methods have looked at this: (1) Bock's nominal model (Bock, 1972; Bock, 1997) provides a statistical description of test takers' qualitative behavior and can be used to make inferences about test taker misconceptions if reasonable distractors are used (ones that seem correct if you have particular misconceptions). In this case, the observations are the response options that test takers selected (not just right/wrong). (2) In cognitive error analysis (Tatsuoka, 1990), an attempt is made to determine what test takers were thinking based on what errors they made. In this case, the observation is the pattern of errors. The content of the items is an integral part of the analysis. (3) Think-aloud protocols can be used to determine how test takers approach the task (e.g., Farr, Pritchard, & Smitten, 1990; Kucan & Beck, 1996; Whitney & Budd, 1996). In this case, the observation is the transcript of what the test taker said. (4) Another method that can be used to explore test taker misconceptions uses test taker notes as an observation. What test takers wrote down while working on the item may give clues about what they were thinking as they worked on the item.

The second example involves inferences about test taker skill profiles. In this case, specific patterns of accuracy on items that are cognitively multidimensional are observed. These patterns are used to determine the test taker's profile of skills on each cognitive domain. This information can be used to provide subscores, score profiles, and diagnostic feedback. Tatsuoka (1990) and Sheehan (1997) have done research in this area. For these models, item content of the particular items answered correctly does matter (unlike in CTT or IRT).

The third example involves inferences about speededness: did test takers have enough time to consider and respond to every item? The observation that has traditionally been used is the number of unanswered items at the end of the test (e.g., Swineford, 1956). When there is no penalty for guessing, test takers often respond randomly to items rather than leave them blank, so different methods are needed to detect speededness on such tests. One approach is to determine if the pattern of incorrect responses at the end of the test is inconsistent with the test taker's performance on the beginning of the test (e.g., Davey, 1990; Yamamoto & Everson, 1995).

All of these examples have involved data that can be collected from paper-and-pencil test administrations. Although these examples do not focus strictly on accuracy, they are based on the response selected, so they do not differ too greatly from accuracy-based approaches. There are other test taker behaviors that could be useful to observe and may require greater shifts in perspective.

The focus of this conference, and the future in assessment in general, is computer-based testing. When computers administer tests, we can record not only the test taker's response, but also how long the person took to make that response (i.e., response times[1]). Response times provide a very different kind of information about test taker behavior and can be used to address various issues, as will be discussed in subsequent sections.

Response Times

Interest in response times as a method of revealing information about mental activity is as old as the field of psychology itself. In 1868, F. C. Donders suggested that the time required to perform a particular mental activity could be inferred by exposing the subject to two procedures that differed only in whether that activity was used

[1] We use the term "response time" rather than "response latency" because the latter term has generally been reserved by cognitive psychologists for unobservable processes (e.g., the duration of the decision component versus the motor component in a simple cognitive task). Response time is used for events with observable start and stop times. We prefer to use the terms with the more established meanings from the field of cognitive psychology.

BEST COPY AVAILABLE

and subtracting the response times (cited in Luce, 1986). The first psychological laboratory was established 11 years later by Wilhelm Wundt, thus marking the official beginning of the field of psychology (Kendler, 1987). Eleven years after that (1890), Joseph Jastrow stated that if the mind is highly structured, then different paths through that structure will require different amounts of time (which will be reflected in response times), and this is one major argument for investigating response times (cited in Luce, 1986).

Response times have been the preferred dependent variable in cognitive psychology since the mid 1950s (Luce, 1986) because how long it takes someone to process something is thought to indicate something about *how* the person processed it. Although response times have had a long history in experimental cognitive psychology, their use in testing has been limited because of the difficulties in recording response times in operational settings. Early research on response times in testing relied on cumbersome methods such as having test takers write down their start and stop times for each item or their total time per item as provided by the examiner. Not only are these intrusive, but these techniques require non-standard test administrations which may preclude inference to typical testing situations (Rindler, 1979).

Now that tests are being administered by computer, response times can be collected easily and unobtrusively in standard, operational settings. As response times are becoming more available, we are in the position to start making use of them. One could even argue that we are obligated to investigate response times in the interest of fairness/equity. The rest of this paper will focus on what response times tell us about test taker behavior and scoring that behavior. Throughout, we will continue to consider what inferences are being made and, in varying degrees, what the implications for scoring might be. In all cases, the observation will be some form of response time (sometimes in combination with accuracy).

Review of Research and Developments to Date

Response times have had a long, but limited, role in testing. For instance, Blommers and Lindquist (1944) investigated the relationship between reading rate and comprehension. Subjects read a question and a paragraph that contained the answer, then answered the question. There was a small positive correlation between the "rate of reading comprehension" (mean reading rate on items answered correctly) and power of comprehension (number-right score). Blommers and Lindquist also found that good comprehenders read more slowly as the difficulty of the material increased, whereas poor comprehenders read at the same rate regardless of difficulty.

Tate (1948) investigated the relationship between speed and accuracy on arithmetic reasoning, number series, sentence completion, and spatial relations questions. Tate found that the fastest test takers were not necessarily the most accurate and that when accuracy is controlled, fast subjects are always fast and slow subjects are always slow. This supports Kennedy's (1930) finding that individuals tend to perform at consistent rates of work across a variety of cognitive tasks, even after partialing out differences due to intelligence. Kennedy concluded that rate of work (speed) is a personality trait.

The use of item response times in testing has been limited due to the difficulty of obtaining responses times in standard testing situations. With the advent of computer-based testing (CBT), however, response times can be collected easily in regular, operational tests. Thus the use of response times has grown substantially in recent years.

It is important to note the difficulty of using response times obtained from complex tasks such as test items. Many researchers have had only limited success with response times in testing (e.g., Bhola, 1994, Kingsbury, Zara, & Houser, 1993, 1994; Parshall, Mittelholtz, & Miller, 1994), clearly indicating that we are still developing methodologies for using response times. It is difficult to determine the best way to use response times in testing and which methods might work.

We will organize research on response times in testing into the following categories: scoring models, speed-accuracy relationships, strategy usage, speededness, pacing, predicting finishing times/setting time limits, and subgroup differences. These categories are closely related to one another. For example, whether response times should be used for scoring will depend on the relationship between speed and accuracy (as well as whether speed is a valid measure of the construct being assessed). Strategy usage (and a special case—speededness) will affect pacing and be affected by time limits. Additionally, all of these issues could vary by subgroup. The

following is not intended to be an exhaustive literature review, but rather the goal is to provide an overview of the type of research that has been done.

Scoring Models

The most direct connection between response-time work and scoring is in the form of response-time models. A number of models have been proposed, and they differ in terms of the assumed response-time distributions (if any), the assumed relationship between ability and response speed, and the nature of items for which the model was designed.

Samejima (1973, 1974, 1983b) offered an extension of IRT that can serve as a model of any continuous response, including response time. Samejima's approach can accommodate the specification of a function relating response time to latent ability; for response times, the function could take the form of a cumulative distribution function. The approach is not limited to a particular function, and Samejima argues that model checking is of utmost importance in function selection. Thus, Samejima's model could, but need not, be used with any of the distributions that have been proposed as appropriate models of response-time distributions.[2] The continuous-response generalization of IRT can also be used with a non-parametric (empirical) approximation of the function relating response time to latent ability.

The use of Samejima's continuous-response generalization of IRT requires that response speed be the dependent variable of interest. Samejima argues that this is appropriate only when the items represent relatively simple cognitive tasks, for which response time is representative of ability. Samejima suggests that response times for more complicated tasks would require more complicated modeling approaches because the response time will have a less straightforward relationship to the cognitive process of interest.

Scheiblechner (1979, 1985) offered the linear exponential model (LEM) as a response-time model for speed test items. He primarily envisioned relatively uncomplicated cognitive tasks such as Posner's perceptual-matching task (e.g., Posner & Boies, 1971) or Sternberg's (1970) memory-scanning task. For the uncomplicated cognitive tasks Scheiblechner describes, response time may be considered a fairly straightforward index of processing. For such tasks, all test takers could probably correctly respond to each item given sufficient time, so errors are likely to be caused by time urgency rather than item difficulty as defined by IRT.

According to the linear exponential model, the response time t for individual i responding to item j follows an exponential distribution with density given by

$$f(t) = (\tau_i + \xi_j)\exp[-(\tau_i + \xi_j)t],$$

where τ_i is a test taker speed parameter and ξ_j is an item speed parameter. The item speed parameter is further modeled in terms of the component processes that are required to solve the item. Thus,

$$\xi_j = \sum_l a_{jl}\eta_l ,$$

where a_{jl} indicates the amount to which component l is present in item j, and η_l indicates the speed of component l.

Scheiblechner focused on items for which response time is often considered an appropriate measure of processing skill. Response accuracy is not considered within the model because most items would be solved correctly were there sufficient time. The applicability of Scheiblechner's model to power tests is an issue for future theory and research.

[2] Schnipke and Scrams (1999) evaluated several candidate functions in terms of the quality of fit to empirical response-time distributions. This work is discussed at the end of this section.

BEST COPY AVAILABLE

Tatsuoka and Tatsuoka (1980) offered a response-time model in which a test taker's speed is characterized by a single parameter, τ, and each item's time requirements are characterized in terms of two parameters, c and u. τ is conceptualized as the expected response time for the given test taker over an infinite set of items of the same type as those presented in the test. c and u are the shape and scale parameters of the Weibull distribution. Together, these parameters determine the distribution of expected response times for a particular test taker responding to a particular item. Tatsuoka and Tatsuoka assumed a Weibull distribution on the basis of earlier work (Tatsuoka & Tatsuoka, 1978), so the cumulative distribution function representing the expected response time, t_{ij}, for test taker i responding to item j is given by

$$F(t_{ij}) = 1 - \exp\left[-\left(\frac{\tau_i + (t_{ij} - \bar{t}_i)}{u_j}\right)^{c_j}\right],$$

where $\bar{t}_i$ is the mean response time for test taker i over all items in the set.

Tatsuoka and Tatsuoka were primarily interested in what might be considered a homogeneous set of items (mixed-number subtraction), and their goal was to use response-time information to classify test takers in terms of the different solution strategies employed. Applying their model to a more heterogeneous group of items would require careful consideration of which items should be included as part of a single set for the determination of $\bar{t}_i$. If test items varied considerably in format or requisite skills and strategies, Tatsuoka and Tatsuoka's model might need to be applied separately to homogeneous subsets. Also note that the authors make no explicit claim about the relationship between response speed and response accuracy; instead, they concentrate on differences in response speed due to differences in strategy usage. They do, however, limit themselves to modeling correct responses on the grounds that incorrect responses may involve various misconceptions at various processing stages, so response times may be less indicative of overall strategy.

Thissen's (1983) timed-testing model is unusual in that it is designed to capture response time and response accuracy within a single model. Thissen, building upon the work of Furneaux (1961), offered a model in which response accuracy is characterized in terms of a two-parameter logistic (2PL) IRT model, and the natural logarithm of response time, $\ln(t)$, is characterized as a linear function of test taker and item parameters:

$$\ln(t_{ij}) = \mu + \tau_i + u_j - \rho z_{ij} + \varepsilon_{ij},$$

where μ is a grand mean across items and test takers, τ_i is a slowness parameter for test taker i, u_j is a slowness parameter for item j, ρ represents the log-linear relationship between response time and ability, z_{ij} is the logit from the IRT model, and ε_{ij} is a normally distributed error term. This model assumes a lognormal distribution for the expected response-time distribution for a test taker responding to an item.

Thissen was interested in both response time and response accuracy, and his model is clearly applicable to items typically found on large-scale, standardized achievement and aptitude tests. The model is also interesting in that Thissen allows for a log-linear relationship between response speed and response accuracy in terms of ρ. Most other response-time models offered by test theorists have ignored issues of response accuracy by focusing on correctly answered items (e.g., Tatsuoka and Tatsuoka, 1980) or on relatively uncomplicated cognitive tasks (e.g., Samejima, 1973, 1974, 1983b; Scheiblechner, 1979, 1985). Thissen makes the interesting point that his model is meant to provide a practical description of responses to typical test items rather than an explanation of the cognitive processes underlying these responses. He suggests that future theorists might develop such process models.

Similar to Samejima (1973, 1974, 1983b) and Scheiblechner (1979, 1985), Verhelst, Verstralen, and Jansen (1997) concentrated on speed tests—tests in which all items could be answered correctly given sufficient time, but time limits lead to a sense of urgency and possible errors. Ability, θ, is tied directly to response speed by introducing the concept of *momentary ability*. Rather than assuming that test takers have fixed ability levels, Verhelst et al. assume that a test taker's momentary (effective) ability is the realization of a random variable

which is influenced both by the test taker's *mental power* and by the time devoted to the item. An item is answered correctly if the test taker's momentary ability exceeds the item's difficulty (conceptualized as a fixed *cognitive weight*). Because the momentary-ability distribution belongs to a shift family in which the location parameter is an increasing function of time, spending more time increases the probability of a correct response.

Verhelst et al. assume that test takers select the amount of time to devote to an item, and these response times follow a two-parameter gamma distribution with rate considered a test taker parameter and shape considered an item parameter. They further assume that the momentary-ability distribution conditioned on response time follows a generalized extreme-value distribution with parameters reflecting the test taker's mental power and the item's cognitive weight. Together, these assumptions determine that the marginal distribution of momentary ability (with time integrated out) follows a generalized logistic distribution. This makes the model consistent with the prevalent use of logistic item response functions. When the shape parameter of the response-time distribution is equal to one, the distribution reduces to the exponential, and the marginal momentary-ability distribution is a one-parameter logistic distribution, so the model is a version of the Rasch (1960), or one-parameter logistic (1PL), IRT model.

The model proposed by Verhelst et al. is similar to Thissen's (1983) model in that response accuracy and response speed are considered simultaneously, but the nature of the speed-accuracy relationship takes a different form in the two models. An important distinction between the two models is the type of items most appropriate for their application. Thissen was primarily interested in items from power tests, whereas Verhelst et al. have focused on items from speed tests. Verhelst et al.'s model may be applied to items from power tests, but this would likely require some modification of the assumed speed-accuracy tradeoff mechanism. Currently, momentary ability (and hence the probability of a correct response) increases without bounds as a function of time. This is clearly an undesirable conceptualization for power tests. Limits could be placed on the amount of time spent, but this is a theoretically unsatisfying solution.

Roskam's (1997) Rasch-Weibull model shares much in common with the model proposed by Verhelst et al. Both models result in Rasch models of response accuracy, but Roskam's model introduces the Rasch response-accuracy model at the level of the correct-response probability conditioned on response time whereas Verhelst et al. introduce the Rasch response-accuracy model for the marginal correct-response probability (integrating over time). Another key difference concerns the assumed response-time distribution—Verhelst et al. assumed gamma-distributed response times, and Roskam assumed Weibull distributions. Both models make use of an effective ability that is an increasing function of time spent, so they explicitly model the relationship between response speed and response accuracy.

Summary. Many response-time models have been proposed, and the present work describes only those that have received considerable attention in the literature. Others include models by Thurstone (1937), White (1973, 1982), and Maris (1993). The models differ in three key areas: the item types most appropriate for the models, the assumed response-time distribution, and the assumed relationship between response speed and response accuracy.

There is little to say about the item types modeled other than to emphasize that most models have focused on relatively uncomplicated cognitive tasks. Exceptions are Tatsuoka and Tatsuoka's (1980) Weibull-based model and Thissen's (1983) timed-testing model. The models offered by Verhelst et al. (1997) and Roskam (1997) may also be applicable to items from power tests, but some modifications would likely be needed.

Although several theorists have tested the appropriateness of their assumed response-time distributions, there have been very few independent investigations along these lines. One notable exception is the work by Schnipke and Scrams (1999) in which they explored the empirical fit of four theoretical response-time distributions: gamma, Gaussian, lognormal, and Weibull. Response times for each of 30 items from a computer-adaptive arithmetic-reasoning test were divided into exploratory (500 observations) and confirmatory (507 to 6,917 observations) samples. Each candidate distribution was fit to each of the exploratory samples by means of maximum-likelihood estimation (Gaussian, lognormal), matching-moments estimation (gamma), or nonlinear regression of the cumulative distribution function (Weibull). The exploratory parameters were used to check each model's fit against the confirmatory samples. Visual inspection of double-probability plots and measures of model misfit showed a strong advantage for the lognormal distribution. This is the distribution proposed by

Thissen (1983). Note that this distribution could also be used in conjunction with Samejima's (1973, 1974, 1983b) continuous-response model.

Interestingly, no validity studies have been offered that address the utility of the resulting scores. Future work on this topic is recommended. Use of the scoring models relies on the relationship between speed and accuracy, and details of empirical work along these lines are provided in the following section.

Speed-Accuracy Relationships

An integral question raised by response-time models concerns the relationship between response speed and response accuracy. This has been an important issue raised by cognitive psychologists as well (Luce, 1986; Townsend & Ashby, 1983), but the focus of cognitive psychologists has been different from the focus of psychometricians. Cognitive psychologists have tended to focus on the within-person relationship between speed and accuracy (the oft-mentioned speed-accuracy tradeoff)—when a person chooses to perform a task more quickly, the person's accuracy tends to decline. Psychometric researchers have tended to focus more on the across-person relationship between speed and accuracy—do the most accurate test takers tend to respond slower or faster than their less accurate counterparts?

For clarification, both types of speed-accuracy relationships can be considered within the response-time model suggested by Verhelst, Verstralen, and Jansen (1997). Speed-accuracy tradeoff is modeled by making momentary ability partially dependent on the time devoted to the task—spending more time on an item increases the probability of a correct response. However, there are also separate parameters reflecting the test taker's mental power (the primary driving force behind accuracy) and mental speed (the rate parameter of the gamma distribution of response times). Any relationship between mental power and mental speed (across test takers) reflects the traditional psychometric focus for speed-accuracy relationships. This type of speed-accuracy relationship is modeled explicitly in Thissen's (1983) timed-testing model by the coefficient relating the logarithm of response time to the IRT logit.

Early in the history of aptitude and achievement testing, researchers believed that speed and accuracy measured the same construct (e.g., Spearman, 1927), implying that it does not matter whether a test taker's ability is measured on a scale of accuracy (power), a scale of speed, or some combination of the two. Starting with the Army Alpha in World War I, the use of time-limited tests has steadily increased. Such tests are often intended to be power tests, but are given with time limits mainly for administrative convenience (Morrison, 1960). If speed and accuracy measure the same construct as Spearman and others believed, time limits will not matter.

Researchers have found, however, that speed and accuracy on complex tasks do not measure the same construct.[3] For example, speed (measured by the time to finish a test) is uncorrelated with test score on an untimed test (Baxter, 1941; Bridges, 1985; Foos, 1989), and time-limit scores (number correct after a given amount of time on a test) are comprised of both speed and level factors (Davidson & Carroll, 1945). Speed (measured by the percentage of people who answered the last item on each page of the test) and accuracy (number right) comprise orthogonal factors in test scores (Myers, 1952). With the advent of computerized testing, the across-person relationship between speed and accuracy has been investigated by a growing number of researchers.

Thissen (1983) investigated his timed-testing model with data from 78 test takers who completed three cognitive tests: a verbal analogies test described by Whitely (1977b) and Tinsley (1971), a subset of Raven's (1956) Progressive Matrices, and part of a spatial-visualization test from the Guilford-Zimmerman Aptitude Survey, Part VI (Guilford & Zimmerman, 1953). Thissen's model allows for two ways to examine the speed-accuracy relationship. First, the relationship is represented by the coefficient relating the logarithm of response time to the IRT logit. This coefficient was low (.20) for the verbal analogies, high (.80) for the progressive matrices, and near-zero (.025) for the spatial-visualization test. A complementary approach is to consider the correlation between the test taker slowness and test taker ability parameters. These values were mediocre for the

[3] Simple reaction times (e.g., Jensen, 1982a, 1982b) have a small negative correlation with simple measures of intelligence, but simple reaction times are not of interest in the present paper and will not be considered further.

verbal analogies (.68), high for the progressive matrices (.94), and near-zero (-.03) for the spatial-visualization test. In all cases, a positive coefficient or correlation indicates that more accurate test takers tended to also be faster.

Segall (1987) investigated the speed-accuracy relationship with data from 209 military applicants completing a pretest of the Computerized Adaptive Test of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). The relationship was investigated in terms of the correlation between total time spent answering items and estimated ability, and analyses were completed separately for nine subtests: general science, arithmetic reasoning, word knowledge, paragraph comprehension, shop information, auto information, math knowledge, mechanical comprehension, and electronic information. The four subtests that involve reasoning or computation (arithmetic reasoning, paragraph comprehension, math knowledge, and mechanical comprehension) demonstrated positive relationships: higher scoring test takers tended to spend more time. The remaining subtests involve simple recall, and they demonstrated either no relationship (auto information, shop information, and electronic information) or a negative relationship (general science, and word knowledge). Note, however, that test takers received different subsets of items. More able test takers received more difficult items, so the observed positive relationships could be explained by more difficult items requiring more time.

Parshall, Mittelholtz, and Miller (1994) investigated a number of test taker, item, and interaction variables as predictors of item response times (transformed onto a logarithmic scale) for an adaptive mathematics placement test. The strongest predictor was the logarithm of the test taker's mean response time across items received, so test takers tended to differ in terms of their response speed. A small amount of variability was explained in terms of estimated ability—more able test takers tended to take slightly more time (replicating Segall's findings). These results suffer from the same confound as did Segall's results—test taker ability is confounded with item difficulty, so the observed relationship may reflect greater time requirements for more difficult items as opposed to faster processing by high-ability test takers.

Bergstrom, Gershon, and Lunz (1994) also attempted to predict item response times (transformed onto a logarithmic scale) from test taker and item characteristics, but unlike Parshall et al. (1994), Bergstrom et al. used a hierarchical linear model with item effects nested within test taker effects. Although they found that relative difficulty was a significant predictor of response time (i.e., test takers spent more time on items that were difficult for them), test taker ability was not a significant predictor of response speed. Bergstrom et al.'s test was considerably different from that used by Parshall et al. (they used an adaptive certification exam instead of an adaptive placement exam), but the results are consistent.

Swanson, Featherman, Case, Luecht, and Nungester (1997) explored the same issue using responses from approximately 20,000 test takers to a computer-administered component of the United States Medical Licensing Exam (USMLE). Consistent with Parshall et al.'s and Bergstrom et al.'s findings, Swanson et al. found no relationship between mean response time and test taker ability.

Scrams and Schnipke (1997) applied a version of Thissen's timed-testing model to responses from approximately 7,000 test takers to three nonadaptive, computer-administered tests assessing verbal-, quantitative-, and analytical-reasoning skills. All three tests resulted in very small values of the coefficient relating the logarithm of response time to the IRT logit: .19, .13, and -.02 for the verbal-, quantitative-, and analytical-reasoning tests, respectively. The correlation between test taker ability and slowness was relatively high for the verbal- and quantitative-reasoning tests (r^2=.39 and .33, respectively) but non-existent for the analytical-reasoning test (r^2=.00). In general, high-scoring test takers tended to take more time than low-scoring test takers on the verbal- and quantitative-reasoning tests, but the groups did not differ on the analytical-reasoning test.

Swygert (1998) investigated the relationship between response time and test taker ability for three subtests of a large-scale, adaptive test. There was no substantive relationship between response time and test taker ability for the verbal-reasoning subtest. There was a modest positive relationship for the quantitative-reasoning subtest, indicating that high-scoring test takers tended to take more time on average. There was also a positive relationship for the analytical-reasoning subtest, but this relationship was due primarily to a small group of outliers. An important difference between Swygert's analyses and others reported in the literature concerns

BEST COPY AVAILABLE

methodology. Swygert analyzed residual effects after controlling for test taker and item differences, so her results do not suffer the confound plaguing the results of earlier CAT work.

Research shows that the relationship between speed and accuracy depends heavily on test context and content. Unfortunately, much of the work addressing this issue uses measures of accuracy that are affected by response speed. The IRT ability estimate, for example, represents the test takers' accuracy given the time constraints of the administration. This is clearly confounded with response speed. This is a serious confound, and a solution to this problem is needed. Some of the response-time scoring models may be useful in this regard, but model-checking procedures will be of considerable importance.

Work on speed-accuracy relationships has implicitly assumed that all test takers use the same strategy for responding to items. Thus, only one speed-accuracy relationship is discussed. If test takers use different strategies, however, the speed-accuracy relationship would depend on strategy. The next section considers strategy usage.

Strategy Usage

Classical test theory (CTT) and item response theory (IRT) describe test takers in terms of their tendency to answer items correctly. CTT and IRT measure quantitative differences in proficiency, but neither is able to detect qualitative differences. CTT and IRT proficiency estimates indicate test takers' ability within a single strategy (Mislevy & Verhelst, 1990).

CTT and IRT characterize ability in terms of the amount or quantity of expertise one has in some domain. CTT and unidimensional IRT characterize expertise in the domain on a single, unidimensional scale. Multidimensional IRT characterizes expertise in the domain on multiple scales. None of these models is appropriate for qualitative changes in ability. Recent research in cognitive and educational psychology indicates that learning is not simply the collection of new knowledge. Rather, learning is an active, constructive process in which learners reconfigure their knowledge structures, incorporate new facts based on their current level of understanding, and create their own interpretations of the material (see Mislevy, 1993; Masters & Mislevy, 1993). For example, expert-novice research suggests that novices use inappropriate or inefficient strategies, based on their internal knowledge structures that they have constructed (see Masters & Mislevy, 1993). Thus we would expect qualitative differences in the performance of experts and novices, or more generally among test takers using different strategies. More important than knowing how many items test takers answer correctly is knowing what strategies test takers are using.

Another example of an important change in strategy is in children's expressive language. When children first learn to speak, they learn words as individual units. Later, as they learn the rules of language, they apply the rules in all cases, even when there are exceptions. Therefore a child might switch from saying "I went to the store" to "I goed to the store" when they learn the "add –ed to a verb to make it past tense" rule. This signifies a deeper understanding of the language, even though they sound like they are regressing. Assessing the child's expressive language in terms of accuracy would result in incorrectly inferring a loss of skill. Incorporating a deeper understanding of strategies would help to uncover the child's improvement.

In CTT and IRT the person either gets the problem right or wrong, but how the person solved the problem is ignored. To take strategy into account, new models are needed. Although models for detecting strategy usage that do not use response times have been proposed (e.g., Davey, 1990; Mislevy, 1996; Samejima, 1983a; Yamamoto & Everson, 1995), the focus here will be on models that do make use of response times to detect strategy usage.

Mislevy and Verhelst (1990) provided a mixture model that takes strategy into account. Their approach characterizes students in terms of strategy used and proficiency within that strategy. Mislevy, Wingersky, Irvine, and Dann (1991) used what has traditionally been called a mental rotation task to illustrate the mixture model approach that was presented in Mislevy and Verhelst (1990). Based on the visualization literature, the mental rotation task can either be solved by mental rotation or by an analytical, rule-based strategy. Different patterns of response times are obtained depending on which strategy is used. These response-time patterns are used to determine which strategy the person most likely used.

As noted in the section on scoring models, Tatsuoka and Tatsuoka's (1980) model was designed to use response-time information to classify test takers in terms of which strategy they employed. Tatsuoka and Tatsuoka (1978) discovered that response-time distributions could be well approximated as Weibull distributions if response times were fit separately for test takers using different subtraction strategies. Tatsuoka and Tatsuoka (1980) built on this finding in developing their response-time model. Test taker strategy is predicted on the basis of model-provided response-time parameters.

Holden (1993) used response times to differentiate between a lying and an honest strategy on personnel tests. Some test takers were instructed to lie and some were instructed to answer honestly. Based on schema theory, it was predicted that test takers would take relatively longer to admit to negative or delinquent behavior if they were lying than if they were answering honestly (e.g., Holden & Kroner, 1992; Holden, Kroner, Fekken, & Popham, 1992). Holden used discriminant function analysis and was able to distinguish between response times for test takers who lied and test takers who answered honestly with a classification hit rate of 64%.

Schnipke (1995) used response times to detect two response strategies on a computer-delivered test that was speeded. In "solution behavior," test takers actively try to determine the correct answer to every item; accuracy is determined by item and test taker characteristics. In contrast, in "rapid-guessing behavior," test takers respond rapidly to items as time expires; accuracy is at or near chance because test takers are not fully considering the items (Schnipke, 1995). Schnipke and Scrams (1997) developed a mixture model to distinguish between these two response strategies on individual items. The model only used response times, but response accuracy supported the classifications. Accuracy was at or below chance for responses classified as rapid guesses. Their model could be applied to other strategy-usage issues as long as the strategies are associated with different response-time demands. For example, "stolen" items might have fast, correct responses.

One special case of strategy usage that has received considerable attention is speededness (e.g., Schnipke, 1995; Schnipke & Scrams, 1997). The next section will discuss the use of response times to define/measure speededness in more depth.

Speededness

In testing, a distinction is made between "power" tests and "speed" tests (e. g., Gulliksen, 1950/1987). In a pure power test, the items range in difficulty and there is no time limit. The goal is to measure how accurately test takers can answer the items. In a pure speed test, the items are very easy and the time limit is very strict. The goal is to measure how quickly test takers can answer items. In reality, most tests contain both speed and power components, and these tests are called speeded (or partially speeded) tests. Speeded tests usually result from administering a power test with a time limit, a practice that is usually required when the test is group-administered.

Speededness, or the extent to which a test is speeded, has traditionally been measured by the percentage of unreached (unanswered) items at the end of the test under the assumption that test takers did not have time to reach these items. Indeed, on speeded tests, some test takers do not finish all of the items. Other test takers, however, rapidly mark answers as time expires, presumably in the hopes of getting some of the items right by chance. Test takers who rapidly mark answers will not have unreached items and therefore are not considered speeded in traditional measure of speededness. Thus traditional speededness indices almost surely underestimate the amount of speededness on most multiple-choice tests (Schnipke, 1995). A more complete estimate of speededness would include test takers who rapidly respond to items as time expires. Thus response times provide additional information regarding speededness.

Some work has used average item response times to investigate speededness. For example, Schaeffer, Reese, Steffen, McKinley, and Mills (1993) compared the average item response time for each item of the computer-based Graduate Record Examination (GRE) General Test (the GRE-CBT) with the amount of time expected if test takers spent the same amount of time on all items and answered all items in the allotted time (i.e., the total amount of time divided by the number of items on the test). Although the average response times were slightly less than the expected times, fewer test takers answered the last item in the quantitative and analytical sections on the GRE-CBT than on the paper-and-pencil version of the GRE. It was concluded that the time limits were

sufficient, but "further insights into the extent of this speededness question may be gained with future examination of the item timing data" (p. 46).

Schnipke (1995) used item response times to explicitly look at speededness. She used accuracy and response time data from a nonadaptive computer-administered test to look at speededness at the end of each test section on the test. She found evidence of two types of speeded behavior. Some test takers did not finish the test. Such test takers spent more time on each item than most test takers and ran out of time before finishing. Figure 1 shows the behavior of such a test taker from Schnipke (1995). The test taker's standardized natural logarithm of response time[4] ($z_{ln(RT)}$) is shown in Figure 1 for each item on the computer-administered linear (nonadaptive) test.

The test taker whose behavior is depicted in FIGURE 1 responded more slowly than most test takers on the first few items (Items 1-7; 1.2 to 2.9 standard deviations above the mean, with a mean of 3 minutes, 10 seconds, per item). The test taker then sped up and responded at about the same speed as most test takers on the next several items (Items 8-13; 0 to 1 standard deviation above the mean, with a mean of 1 minute, 34 seconds, per item). The test taker did not finish the rest of the test. This test taker was slow, but accurate: of Items 1-12, all were answered correctly except Item 9. The test taker did not respond to Item 13 (i.e., it was omitted), and the rest (Items 14-25) were unreached (i.e., the items were never displayed on the screen).

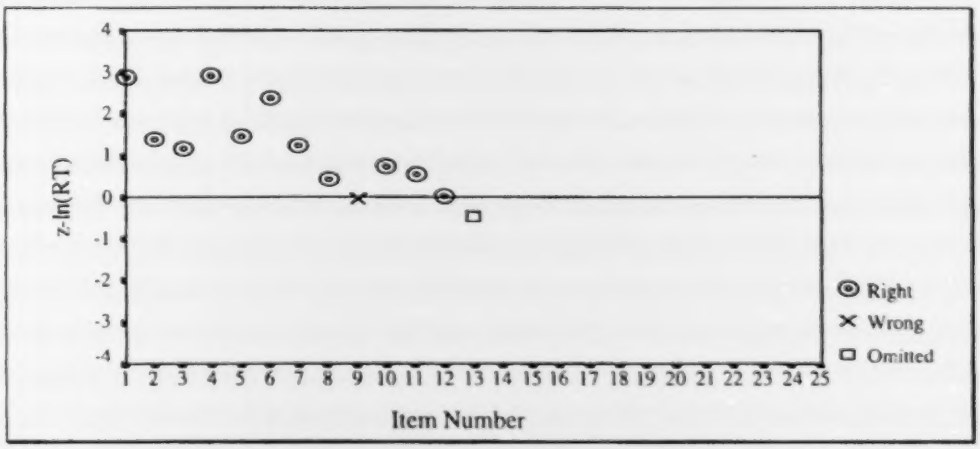

FIGURE 1. *Standardized transformed response time, $z_{ln(RT)}$, across items for a speeded test taker who did not finish the test*

The other type of speeded behavior that Schnipke (1995) found was rapid-guessing behavior (defined in the previous section on strategy usage). Figure 2, from Schnipke (1995), shows the behavior of a test taker who switched from solution behavior to rapid-guessing behavior. At the beginning of the test, the test taker responded at a slower rate than most test takers (from .5 to 1.5 standard deviations above the mean, with a mean of 2 minutes and 17 seconds, per item). After Item 13 (when the test taker had only two minutes remaining), the test taker suddenly started responding much faster (from -.7 to -3.9 standard deviations below the mean, with a mean 12.25 seconds per item). The test taker's accuracy also changed—from 77% accuracy on the first 13 items to 25% accuracy on the last 12 items. The test taker engaged in solution behavior on the first half of the items, but on the last half of the items (and with very little time left), the test taker engaged in rapid-guessing behavior.

[4] The natural logarithm is used because the response-time distributions are positively skewed; the natural logarithm transformation creates a more normal distribution. The transformed response times were then standardized, item by item, to control for item differences (e.g., long items take longer, on average, to answer than short items). If a test taker responded at the mean speed on an item, $z_{ln(RT)}$ would be 0. Positive values of $z_{ln(RT)}$ indicate that the test taker responded slower than the average of all test takers, and negative values indicate that the test taker responded faster than average.

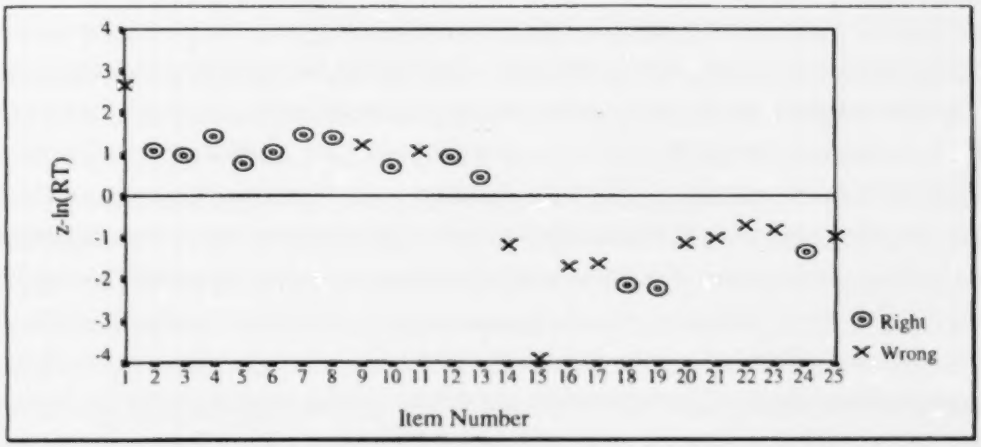

FIGURE 2. *Standardized transformed response time, $z_{ln(RT)}$, across items for a speeded test taker who switched strategies on the second half of the test*

Schnipke and Scrams (1997) developed a mixture model of two lognormal distributions to distinguish between solution behavior and rapid-guessing behavior on a nonadaptive computerized test. The model used response times to determine whether a response was more likely to come from the rapid-guessing distribution or the solution distribution. Schnipke and Scrams determined that the rapid-guessing distribution is essentially the same across items, supporting the claim that rapid-guessing behavior is not affected by item content.

Figure 3 shows two sample items from Schnipke and Scrams (1997). The item on the left (the 4th item in the 25-item nonadaptive test) shows no rapid-guessing behavior, whereas the item on the right (the 17th item in the test) contains both rapid-guessing behavior and solution behavior. Schnipke and Scrams' final mixture model indicated that 1% of the test takers on Item 4 engaged in rapid-guessing behavior, whereas 18% engaged in rapid-guessing behavior on Item 17.

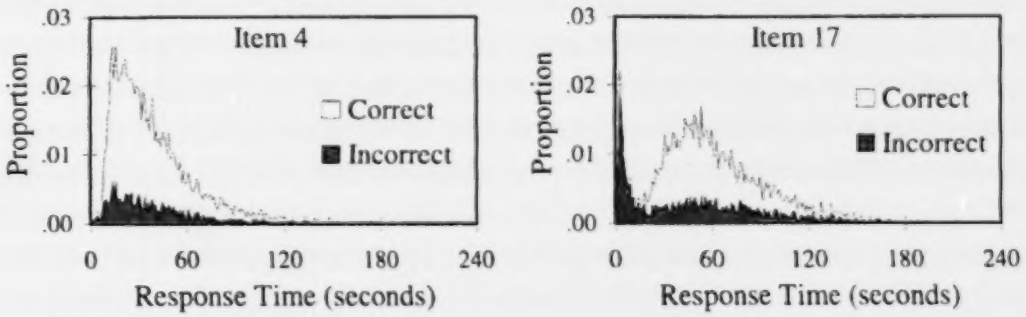

FIGURE 3. *Probability density functions of response times for two items. Item 4 contains only solution-behavior responses. Item 17 contains both solution-behavior responses and rapid-guessing-behavior responses.*

Swanson, Featherman, Case, Luecht, and Nungester (1997) investigated response times on computer-administered components of the USMLE. They investigated both a nonadaptive computer-based test (CBT) and a testlet-based computerized adaptive sequential test (CAST). CAST adapts between testlets, but not within

BEST COPY AVAILABLE

testlets. The testlets used by Swanson et al. contained 60 items. Both the CBT and CAST showed evidence of rapid-guessing behavior, as shown in response-time density functions (similar to Figure 3, Item 17, above). Swanson et al. determined that the CBT was speeded for less proficient test takers. The CAST design reduced speededness for less proficient test takers, but it may have increased speededness for more proficient test takers.

Hadadi and Luecht (1998) analyzed data from two CBTs – one that was speeded and one that was not speeded. They separated their data into blocks of 30 items (Positions 1-30, 31-60, etc.) and combined response times from those blocks. This allowed them to investigate stable response-time distributions to detect rapid-guessing behavior. In particular, the speeded CBT showed evidence of both omits (unanswered items) and rapid-guessing behavior (identified as response times of less than 8 seconds), especially on the last 30 items (out of 180 items).

Assessing speededness in CATs is made difficult because test takers do not see the same items in the same positions and because individual items might not be seen by enough test takers to draw meaningful conclusions about response times. To look for speededness at the end of the test, results must be pooled over different items or must be done at the test taker level. Hadadi and Luecht's (1998) is one approach. Swygert's (1998) solution for investigating response times by item position on a CAT is to remove the item and test taker effects from each item and to look at the residual response times across items for each position. Bontempo and Julian (1997) used a similar procedure for investigating speededness on a variable-length adaptive test (the NCLEX-RN™ examination). All test takers were required to take a minimum of 60 scored items (and 15 unscored items). If a pass/fail decision could be made at that point, testing stopped. Otherwise test takers continued to take items, up to a total of 265 (250 scored and 15 unscored), until a reliable pass/fail decision could be made. Bontempo and Julian assumed that the first 60 items would be completely unspeeded. One way they assessed speededness was by comparing test takers' relative response rate on the first 60 scored items to their relative response rate on the remaining 190 items or the last 50 items. Relative response rate, τ_i, is given by

$$\tau_i = \frac{\sum_{j=1}^{J} \left(t_{ij} - \bar{t}_j \right)}{J},$$

where t_{ij} is the response time for test taker i to item j, $\bar{t}_j$ is the average response time on item j for all test takers who took the item, and J is the total number of items in the comparison. Bontempo and Julian found that 88% of the test takers worked faster on the remaining 190 items, and 91% worked faster on the last 50 items, indicating that test takers had to work faster at the end of the test, indicating that it may have been speeded. Overall, Bontempo and Julian concluded that at least 5% of the test takers were significantly affected by the time limit. However, Bontempo and Julian did not establish that test taker accuracy was affected by increased speed, so their results may reflect increased efficiency or comfort with the delivery system.

Rapid-guessing behavior and not finishing a speeded test are two pacing strategies. Many other pacing strategies are possible, and the next section will discuss test taker pacing more generally.

Pacing

The goal of standardized testing is to assess test taker performance, and this is usually based on the accuracy with which test takers respond to items under the assumption that more able test takers will tend to respond more accurately. Unfortunately, accuracy can be affected by factors other than ability, such as strategy usage, processing speed, or test sophistication. Test taker pacing is related to all three of these: pacing may represent strategic choices, test taker processing speed may influence pacing decisions, and more test-sophisticated test takers may choose more optimal pacing strategies than less test-sophisticated test takers of comparable ability. For the purpose of the present work, pacing issues will be limited to response-time patterns across item positions, types, or difficulties. Issues related only to overall test taker speed are discussed in the sections on Speed-Accuracy Relationships, Subgroup Differences, and Speededness.

Llabre and Froman (1987) investigated the pacing strategies of 38 Hispanic and 28 non-Hispanic, white test takers who completed a computer-administered version of the California Test of Mental Maturity. Although Hispanic test takers tended to spend slightly more time on each item, the more interesting finding is that non-Hispanic test takers tended to allocate their testing time according to item difficulty to a greater extent than did Hispanic test takers. Llabre and Froman argued that this was a more optimal strategy, and suggested that non-Hispanic test takers may have been more test sophisticated than ability-matched Hispanic test takers.

Gitomer, Curtis, Glaser, and Lensky (1987) offered one of the most interesting uses of response times for psychometric tests. They investigated processing strategies for analogy problem solving by examining timed eye-fixation patterns observed with eye-tracking equipment. One of many interesting results was the observation that high- and low-ability test takers tended to allocate their processing time similarly on easy analogies, but high-ability test takers tended to increase processing time for more difficult items to a greater extent than did low-ability test takers. This is consistent with Llabre and Froman's (1987) findings. Unlike Llabre and Froman (1987), Gitomer et al. were able to determine that the increased processing time was for both stem and response-option processing.

Schaeffer, Reese, Steffen, McKinley, and Mills (1993) investigated a number of response-time issues using data from a field test of the computer-administered (nonadaptive) GRE General Test. They found that although low-ability test takers tended to spend similar amounts of time on the easiest and most difficult items administered, middle-ability test takers spent more time on difficult items, and high-ability test takers tended to spend considerably more time on difficult items. This essential pattern was observed for all three subtests: verbal-, quantitative-, and analytical-reasoning. Unfortunately, item type was confounded with item difficulty, and the middle-difficulty items failed to fit the expected pattern for verbal- or quantitative-reasoning items. Examination of the mean response times by item type and test taker ability revealed that low- and high-ability test takers tended to allocate their testing time differently across item types. For some item types, high-ability test takers invested less time than low-ability test takers; the opposite pattern was observed for other item types. Some item types showed no response-time differences across ability groups.

Swanson, Featherman, Case, Luecht, and Nungester (1997) also investigated pacing differences across ability groups. Similar to Schaeffer et al.'s (1993) results, Swanson et al. found that low-ability test takers tended to allocate their time equally across items of varying difficulty. High-ability test takers, on the other hand, tended to allocate more time to difficult items and less time to easy items. Swanson et al. also observed that test takers tended to speed up throughout the test—mean item response times tended to decrease across sequential 30-item test blocks. This latter finding is consistent with results reported by Bontempo and Julian (1997) for a large-scale, adaptive mastery test. They found that test takers tended to respond more slowly to the first 60 items than to the remaining items.

Schnipke (1995) explored pacing strategies at the test taker level for a computer-administered, nonadaptive test. She suggested that plots of standardized response times against item position could be useful for investigating pacing issues. Her primary interest was in identifying test takers who engaged in speeded behavior. To this end, she provided examples of test takers who tended to respond more slowly than most test takers during the early part of a test and then either failed to finish (see Figure 1) or responded very rapidly to later items (see Figure 2).

More recent work by Scrams and Schnipke (1999) has used the same technique as Schnipke (1995) with data from an adaptive test to uncover additional pacing strategies including equal-allocation strategies and decreasing-speed strategies. The latter strategy was unexpected, but Scrams and Schnipke found a large minority of test takers who tended to respond quickly to early items and were then able to spend more time on later items. The advantages of such a strategy in a CAT environment should be considered in future research.

Predicting Finishing Times/Setting Time Limits

This section will cover studies that have used item response times in attempt to determine/predict finishing times or set time limits for computerized tests. Studies that used total test times (rather than item-level response

BEST COPY AVAILABLE

times) are not included (e.g., Segall, Moreno, Kieckhaefer, Vicino, & McBride, 1997), nor are studies that investigated the effects of different time limits on test performance.

Time limits define the task for the test takers (Mollenkopf, 1960), which is another way of saying that it affects test taker behavior. Time limits will also affect the meaning of scores. As Mollenkopf (1960) notes, identical material administered under speeded and power conditions may not be measuring the same thing. It is important to remember this when interpreting scores from speeded tests, especially if the degree of speededness is unclear. The following studies attempt to specify optimal time limits based on item response times (Bhola, Plake, & Roos, 1993), predict response times so that the regression equation can be used to predict finishing times and therefore inform appropriate time limits (Bergstrom, Gershon, & Lunz, 1994; Halkitis, Jones, & Pradhan, 1996; Reese, 1993), and use response-time information to constrain item selection in a CAT (van der Linden, Scrams, & Schnipke, *in press*).

Bhola, Plake, and Roos (1993) administered a computerized test without a time limit and recorded the amount of time test takers spent on each item. They calculated the median and interquartile range of response times for each item. They suggest that the optimum time limit for a nonadaptive test would be the sum of the median response time for each item, plus .5 times the interquartile range. This should allow approximately 75% of the test takers to complete the test under power conditions (and the remaining 25% under speeded conditions). They demonstrate that two test forms matched on difficulty, discrimination, and lower asymptote (from the three-parameter logistic IRT model) can have very different optimal time limits (according to their formula). One of their hypothetical forms required 28% more time than the other, thus suggesting that item-time requirements need to be balanced across test forms along with statistical and content constraints.

For the GRE-CAT, time limits were established initially by building a regression model that predicted response times (from a nonadaptive computer-administered field test version of the GRE) based on test taker ability and item characteristics (Reese, 1993). The model was used with simulated CAT data to predict finishing times on the CAT. Time limits for the CAT were selected to ensure that virtually all test takers were predicted to have enough time to finish the CAT. The time limits were found to be appropriate after actual CAT timing data were collected (Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995).

Halkitis, Jones, and Pradhan (1996) investigated the relationship between response times and item characteristics (item length, difficulty, and discrimination) on a nonadaptive computer-administered test. Item length was defined as the number of words in the item. Item difficulty was defined as the percentage of test takers who answered correctly. Item discrimination was defined as the point biserial correlation coefficient between the item score (right/wrong) and total test score. Halkitis et al. predicted the logarithm of response times from item length, difficulty, and discrimination and found that 50.18% of the variance in log response times could be predicted by these three variables. They suggest that their results provide a preliminary model that can be used to provide an initial estimate of the total testing time required.

Bergstrom, Gershon, and Lunz (1994) investigated the relationship between response times and test taker and item characteristics on a CAT. They used hierarchical linear modeling to predict the logarithm of response times from test taker characteristics (gender, ethnicity, ability, and anxiety) and item characteristics (item position, item length, relative item difficulty, key position, and inclusion of a figure). Gender, ethnicity, and ability did not affect response times, although all other variables did. They were able to explain at least one third of the variation in log response times for most test takers. They suggested that to control total testing time, items with characteristics that make them especially time consuming could be removed from the item bank or balanced across test takers during test administration. Further, they suggest that understanding how item characteristics impact response times may allow test developers to predict the amount of time required for test administration. Such predictions can be based on the response-time history of individual test items.

A model for incorporating response time information into item selection for CAT has been proposed by van der Linden, Scrams, and Schnipke (*in press*). They used a model that describes response times using both item and test taker speed parameters. The test taker speed parameter is updated after every response and is used to guide item selection (based on IRT parameters and item speed) to ensure that all test takers will have enough time to finish the test. They found that the algorithm was able to reduce speededness for test takers who

otherwise would have suffered from the time limit. This algorithm directly controls finishing times for each test taker to make sure that test takers have enough time to work on all items.

Research on setting time limits (or predicting finishing times) has included investigating subgroup differences as time limits should be set, in part, to limit differential subgroup speededness. Additional research on subgroup differences in response times is discussed in the next section.

Subgroup Differences

There is evidence that speed and the concept of time are perceived differently by various cultures (Klineberg, 1935; Roberts & Greene, 1971; Shannon, 1976). Therefore, it is possible that subgroups will have different patterns of response times and may be differentially affected by time limits on tests. Thus, it is relevant to investigate response times for various subgroups of the population, as well as for the overall test taker population.

Differential access to computers has been documented for members of different gender and ethnic subgroups and social classes (Sutton, 1991), although more recent work has not found such differences (Jenkins & Holmes, 1999). If subgroups differ in computer access, low-access groups may be disadvantaged on computer-based tests. Even if certain subgroups do not have less access to computers, their scores could be affected in irrelevant ways because of the computer-delivery of the test. Thus subgroup analyses are especially important in computer-based tests.

Researchers have not found gender and ethnicity to be significant predictors of response times on computer-administered tests (Bergstrom, Gershon, & Lunz, 1994; Parshall, Mittelholtz, 1994; Schnipke, 1995), although anxiety was found to significantly predict response times (Bergstrom et al., 1994).

Other researchers have found small differences in response times for various subgroups. Llabre and Froman (1987) found that Hispanic test takers consistently spent more time on items than white test takers. O'Neill and Powers (1993) found that male and female test takers differed very little on the amount of time spent on tutorials and tests, although African American, Hispanic, and Native America test takers generally spent more time on tutorials and test sections than did white test takers. Schaeffer et al. (1993) found that female test takers spent slightly longer on tutorials than did male test takers, but the differences were small, and both groups saw and answered almost every item. African America test takers, however, spent more time on tutorials than did white test takers, and they also saw and answered one fewer item on average than did white test takers. Schnipke (1995) found that rapid-guessing behavior was more common among male test takers on an analytical test, more common among female test takers on a quantitative test, and equally common on a verbal test. Female test takers were slightly more likely to have unreached items on all three tests.

Most researchers who have investigated subgroup differences in response times have compared median or mean response times. Schnipke and Pashley (1997) suggested a more powerful approach of looking at the entire distribution of response times for each item. They did this using survival analysis methodology, and they demonstrated the approach with CBT data comparing response times for primary and nonprimary speakers of English. They found that nonprimary speakers responded slower on some items at the beginning of the test and faster on some items toward the end of the test, suggesting that the nonprimary speakers were more affected by the speed component than were primary speakers.

Overall, there is evidence of small response-time differences between various subgroups but these small differences tend to be masked by stronger predictors in regression analyses. Such differences could result in differential speededness, differential finishing rates (e.g., O'Neill & Powers, 1993; Schaeffer et al., 1993; Schnipke, 1995), or different test-taking strategies. As noted by O'Neill and Powers (1993) subgroup timing information should be taken into account when time limits are set to ensure that no subgroups are disadvantaged. Examining subgroup differences in response time is not only possible, but it is necessary to ensure equity.

Recommendations for Future Research

In CTT, test takers are ranked in terms of their propensity to perform tasks in a given domain. This propensity is inferred from performance on a subset of domain tasks, and domain sampling is very important. In contrast, in IRT, test takers are ranked in terms of a hypothesized underlying trait. Estimation of this trait is based on established statistical properties of items and an assumed relationship between item properties and the underlying trait. In CTT and 1PL (Rasch) IRT, scores depend only on the number of correct responses; in non-Rasch IRT, the statistical properties of items result in differential weights for scores. Neither CTT nor traditional IRT make explicit use of item content.

Over the past few decades, cognitive perspectives have become dominant in psychology, and several theorists have argued for a similar paradigm shift within psychometrics (Embretson, 1983, 1997; Frederiksen, Mislevy, & Bejar, 1993; Mislevy, 1994, 1996; Snow & Lohman, 1989; Tatsuoka, 1990; Yamamoto, 1989). Response-time researchers should take these recommendations to heart, as response times have been the preferred dependent variable in cognitive psychology since its inception (Luce, 1986; Townsend & Ashby, 1983). Although most of the psychometric response-time models rely at least partially on work from cognitive psychology, little of the empirical work has used either these models or other approaches present in the cognitive psychology literature.

Two favorable exceptions are the work by Whitely (1977a) and Gitomer, Curtis, Glaser, and Lensky (1987). Whitely tested a multiple-component model of analogical reasoning using confirmatory factor analysis. The model was based on theory concerning the cognitive-processing demands of analogies, and Whitely tested the model against both accuracy and response-time data. The model appeared adequate for accuracy data, but was insufficient to explain the response-time data. Whitely argued that a more detailed model of cognition might be needed.

Gitomer et al. (1987) also investigated the processes underlying analogy performance. They used eye-tracking equipment to monitor the amount and proportion of time spent on the initial encoding of the problem stem, subsequent encoding of the stem, and encoding of the response options. Similar to Whitely (1977a), Gitomer et al. were motivated by cognitive theories of analogical reasoning. Interestingly, Gitomer et al. found that both low- and high-ability test takers tended to increase their processing time as items became more difficult, but low-ability test takers tended to increase the time spent on initial encoding of the stem whereas high-ability test takers increased their subsequent processing time. Gitomer et al. inferred that high-ability test takers focus on response options for difficult items whereas low-ability test takers focus on initial encoding of the stem. This suggests a fundamental difference in how the two groups are processing items.

Work of this type is very valuable and makes excellent use of response-time information. Response times have been the preferred dependent variable among cognitive psychologists because they believe that response times provide information about how individuals process information (Luce, 1986; Townsend & Ashby, 1983). Such a perspective would greatly benefit psychometricians who are interested in going beyond the relatively simple scores derived from CTT and IRT, but the nature of such work may demand closer interactions among psychometricians and cognitive psychologists. Fortunately, such interactions are likely to also offer benefits for other issues in psychometrics such as complex assessment, diagnostic feedback, and construct validity.

References

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. *Journal of Educational Psychology, 32*, 285-296.

Bergstrom, B., Gershon, R., & Lunz, M. E. (1994, April). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bhola, D. S. (1994). An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Bhola, D. S., Plake, B. S., & Roos, L. L. (1993, October). *Setting an optimum time limit for a computer-administered test*. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago.

Blommers, P., & Lindquist, E. F. (1944). Rate of comprehension of reading: Its measurement and its relation to comprehension. *Journal of Educational Psychology, 35*(8), 449-473.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York: Springer.

Bontempo, B. D., & Julian, E. R. (1997, March). *Assessing speededness in variable-length computer adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Bridges, K. R., (1985). Test-completion speed: Its relationship to performance on three course-based objective examinations. *Educational and Psychological Measurement, 45*, 29-35.

Davey, T. (1990, April). *Modeling timed test performance*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Davidson, W. M, & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement, 5*, 411-427.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-322). New York: Springer.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*, 209-226.

Foos, P. W. (1989). Completion time and performance on multiple-choice and essay tests. *Bulletin of the Psychonomic Society, 27*, 179-180.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum.

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology* (pp. 167-192). London: Pitman.

Gitomer, D. H., Curtis, M. E., Glaser, R., & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. *Journal of Educational Psychology, 79*, 212-219.

Guilford, J. P., & Zimmerman, W. F. (1953). *The Guilford-Zimmerman aptitude survey. IV. Spatial Visualization Form B*. Beverly Hills, CA: Sheridan Supply Company.

Gulliksen, H. (1950/1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum.

BEST COPY AVAILABLE

Hadadi, A., & Luecht, R. M. (1998, April). *Effects of assessing speededness/non-speededness in computerized tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996, April). *Estimating testing time: The effects of item characteristics on response latency*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Holden, R. R. (1993, August). *Response latency detection of lying on personnel tests*. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 4*, 170-173.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63*, 272-279.

Jenkins, S. M., & Holmes, S. D. (1999). *Computer usage and access patterns of actual and potential LSAT takers*. (Computerized Testing Report 97-10). Newtown, PA: Law School Admission Council.

Jensen, A. R. (1982a). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer.

Jensen, A. R. (1982b). The chronometry of intelligence. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1). London: Lawrence Erlbaum.

Kendler, H. H. (1987). *Historical foundations of modern psychology*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Kennedy, M. (1930). Speed as a personality trait. *Journal of Social Psychology, 1*, 286-298.

Kingsbury, G. G., Zara, A. R., & Houser, R. L. (1993, April). *Procedures for using response latencies to identify unusual test performance in computerized adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Kingsbury, G. G., Houser, R. L., & Zara, A. R. (1994, April). *Modeling item response latencies in computerized adaptive tests*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Klineberg, O. (1935). *Race differences*. New York: Harper.

Kucan, L., & Beck, I. L. (1996). Four fourth graders thinking aloud: An investigation of genre effects. *Journal of Literacy Research, 28*, 259-287.

Llabre, M. M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. *Journal of Experimental Education, 55*, 137-140.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58*, 445-469.

Masters, G. N., & Mislevy, R. J. (1993). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 219-242). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Mislevy, R. J., Wingersky, M. S., Irvine, S. H., & Dann, P. L. (1991). Resolving mixtures of strategies in spatial visualization tasks. *British Journal of Mathematical and Statistical Psychology, 44*, 265-288.

Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement, 20*, 223-230.

Morrison, E. J. (1960). On test variance and the dimensions of the measurement situation. *Educational and Psychological Measurement, 20*, 231-250.

Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17*(3), 347-352.

O'Neill, Kathleen, & Powers, D.E. (1993, April). *The performance of examinee subgroups on a computer-administered test of basic academic skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Parshall, C. G., Mittelholtz, D. & Miller, T. R. (1994, April). Response time: An investigation into determinants of item-level timing. In C. G. Parshall (Chair), *Issues in the development of a computer adaptive placement test*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Posner, M. I., & Boies, S. (1971). Components of attention. *Psychological Review, 78*, 391–408.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raven, J. (1956). *Progressive matrices*. London: Lewis.

Reese, C. M. (1993, April). Establishing time limits for the GRE computer adaptive tests. In W. D. Way (Chair), *Practical problems in the development of large scale computer adaptive tests*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*(4), 261-270.

Roberts, A. H., & Greene, J.E. (1971). Cross-cultural study of relationships among four dimensions of time perspective. *Perceptual and Motor Skills, 33*, 163-173.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer.

Samejima, F. (1973). Homogeneous case of the continuous response level. *Psychometrika, 38*, 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Samejima, F. (1983a). *A latent trait model for differential strategies in cognitive processes* (Technical Report ONR/RR 81-1). Knoxville, TN: University of Tennessee.

Samejima, F. (1983b). *A general model for the homogeneous case of the continuous response* (ONR Research Report 83-3). Arlington, VA: Office of Naval Research. Personnel and Training Research Programs.

Schaeffer, G. A., Reese, C. M., Steffen, M., McKinley, R. L., & Mills, C. N. (1993). *Field test of a computer-based GRE General Test* (Research Report No. 93-07). Princeton, NJ: Educational Testing Service.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE General Test*. (Research Report No. 95-20). Princeton, NJ: Educational Testing Service.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology, 19*, 18–38.

Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 219-244). Orlando, FL: Academic Press.

Schnipke, D. L. (1995). Assessing speededness in computer-based tests using item response times (Doctoral dissertation, Johns Hopkins University, 1995). *Dissertation Abstracts International, 57*, B759.

Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks* (Computerized Testing Report 97-09). Newtown, PA: Law School Admission Council.

Scrams, D. J., & Schnipke, D. L. (1997). *Making use of response times in standardized tests: Are accuracy and speed measuring the same thing?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Scrams, D. J., & Schnipke, D. L. (1999). *Response-time feedback on computer-administered tests*. Unpublished manuscript.

Segall, D. O. (1987). Relation between estimated ability and test time on the CAT-ASVAB. Unpublished manuscript.

Segall, D. O., Moreno, K. E., Kieckhaefer, W. F., Vicino, F. L., & McBride, J. R. (1997). Validation of the experimental CAT-ASVAB system. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 103-114). Washington, DC: American Psychological Association.

Shannon, L. (1976). Age change in time perception in Native Americans, Mexican Americans, and Anglo-Americans. *Journal of Cross-Cultural Psychology, 1*, 117-122.

Sheehan, K. M. (1997). *A tree-based approach to proficiency scaling and diagnostic assessment* (Research Report No. 97-9). Princeton, NJ: Educational Testing Service.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Spearman, C. (1927). *The abilities of man*. New York, Macmillan.

Sternberg, S. (1970). Memory scanning: Mental processes revealed by reaction time experiments. In J. S. Antrobus (Ed.), *Cognition and affect* (pp. 13-58). Boston: Little, Brown.

Sutton, R. E. (1991). Equity and computers in schools: A decade of research. *Review of Educational Research, 61* (4), 475-503.

Swanson, D. B., Featherman, C. M., Case, S. M., Luecht, R., & Nungester, R. (1997, March). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Swineford, F. (1956). *Technical manual for users of test analysis*. (Statistical Report 56-42). Princeton, NJ: Educational Testing Service.

Swygert, K. A. (1998). *An examination of item response times on the GRE-CAT*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement, 8*, 353-374.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K., & Tatsuoka, M. M. (1978). *Time-score analysis in criterion-referenced tests* (CERL Report E-1). Urbana, IL: University of Illinois, Computer-Based Education Research Laboratory.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference* (pp. 236–256). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

BEST COPY AVAILABLE

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2*, 249-254.

Tinsley, H. E. (1971). An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Unpublished doctoral dissertation, University of Minnesota.

Townsend, J. T. & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (*in press*). Using response-time constraints to control for differential speededness in computerized adaptive testing. Conditionally accepted for publication at *Applied Psychological Measurement*.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York: Springer.

White, P. O. (1973). Individual differences in speed, accuracy and persistence: A mathematical model for problem solving. In H. J. Eysenck (Ed.), *The measurement of intelligence* (pp. 246-260). Baltimore: Williams and Wilkins.

White, P. O. (1982). Some major components in general intelligence. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 44–90). Berlin: Springer-Verlag.

Whitely, S. E. (1977a). Information-processing on intelligence test items: Some response components. *Applied Psychological Measurement, 1*, 465-476.

Whitely, S. E. (1977b). Relationships in analogy items: A semantic component of a psychometric task. *Educational and Psychological Measurement, 37*, 725-739.

Whitney, P, & Budd, D. (1996). Think-aloud protocols and the study of comprehension. *Discourse Processes, 21*, 341-351.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models*. (Research Report No. 89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified Hybrid model*. (Research Report No. 95-16). Princeton, NJ: Educational Testing Service.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").